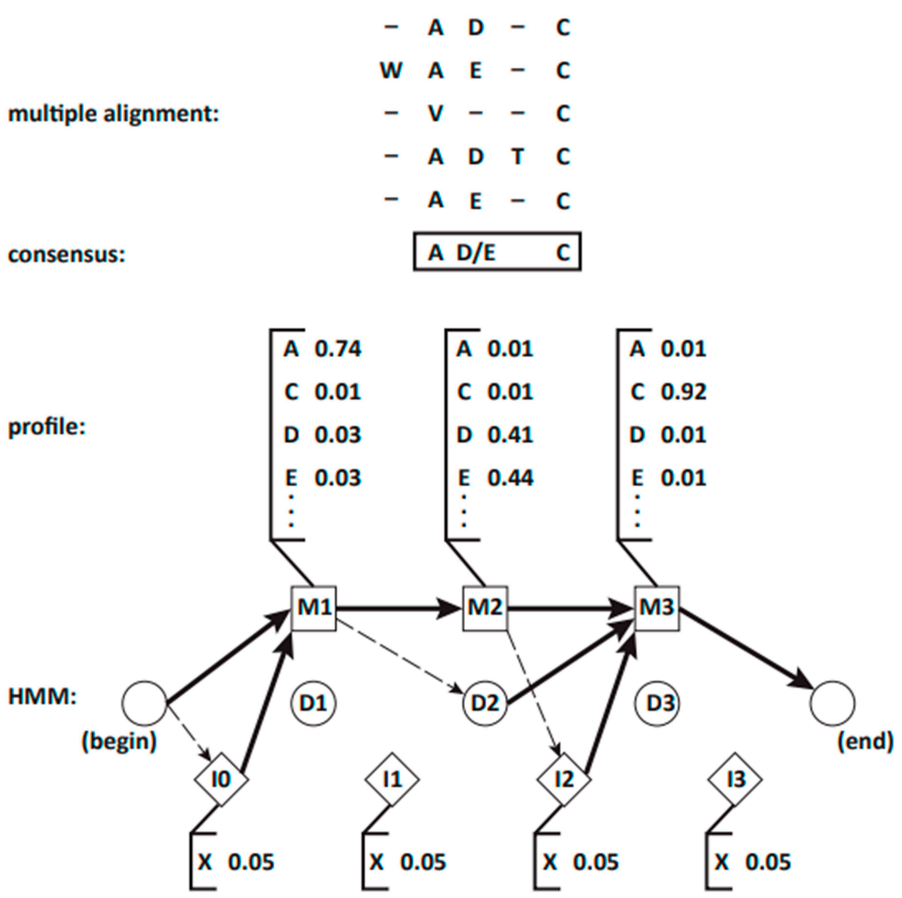
Biological Sequence Analysis using *Profile Hidden Markov Models*

Model Introduction

In the model, each column of symbols in the alignment is represented by a frequency distribution of the symbols (called a "state"), and insertions and deletions are represented by other states. One moves through the model along a particular path from state to state in a Markov chain (i.e., random choice of next move),A trying to match a given sequence. The next matching symbol is chosen from each state, recording its probability (frequency) and also the probability of going to that state from a previous one (the transition probability).



Dataset

The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).

Dataset used: [Family: Miga (PF10265)]
Mitoguardin (Miga) was first identified in flies as a mitochondrial outer-membrane protein that promotes mitochondrial fusion.

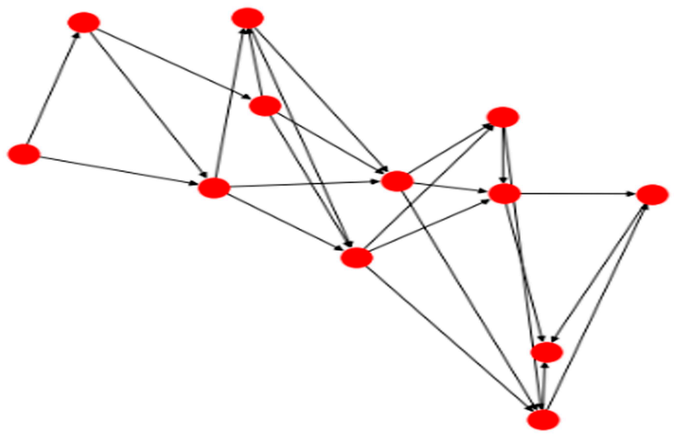
In bioinformatics and biochemistry, the FASTA format is a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences.

Evaluation and Perfomance

The Baum-Welch (Expectation Maximization) training is relatively slow, as it attempts to optimize all possible paths through an HMM for a given observation. In contrast, the Viterbi training algorithm only attempts to optimize the most likely path associated with a sequence, sacrificing accuracy for speed.

Graph construct using Viterbi:

Sequence: 'ACT' -- Log Probability: -0.5132449003570658 -- Path: M1 M2 M3
Sequence: 'GGC' -- Log Probability: -11.048101241343396 -- Path: I0 I0 D1 M2 D3
Sequence: 'GAT' -- Log Probability: -9.125519674022627 -- Path: I0 M1 D2 M3
Sequence: 'ACC' -- Log Probability: -5.0879558788604475 -- Path: M1 M2 M3



Federal University of Minas Gerais
Computer Science Department

Mírian Francielle
Undergraduate Student of Computational Math
mirianfrancielle@dcc.ufmg.br